Some Aspects of Design of Experiments

Nancy Reid
University of Toronto, Toronto Canada

Abstract
This paper provides a brief introduction to some aspects of the theory of design of experiments that may be relevant for high energy physics experiments and associated Monte Carlo investigations.

1 Introduction

'Design of experiments' means something specific in the statistical literature, which is different from its more general use in science. The key notion is that there is an *intervention* applied to a number of *experimental units*; these interventions are conventionally called treatments. The treatments are usually assigned to experimental units using a randomization scheme, and randomization is taken to be a key element in the concept in the study of design of experiments. The goal is then to measure one or more responses of the units, usually with the goal of comparing the responses under the various treatments. Because the intervention is under the control of the experimenter, a designed experiment generally provides a stronger basis for making conclusions on how the treatment affects the response than an observational study.

The original area of application was agriculture, and the main ideas behind design of experiments, including the very important notion of randomization, were developed by Fisher at the Rothamsted Agricultural Station, in the early years of the twentieth century. A typical agricultural example has as experimental units some plots of land, as treatments some type of intervention, such as amount of or type of fertilizer, and as primary response yield of a certain crop. The theory of design of experiments is widely used in industrial and technological settings, where the experimental units may be, for example, manufactured objects of some type, such as silicon wafers, the treatments would be various manufacturing settings, such as temperature of an oven, concentration of an etching acid, and so on, and the response would be some measure of the quality of the resulting object. In so-called *computer experiments*, the experimental units are simulation runs, of, for example, a very complex system such as used for climate modelling or epidemic modelling; the 'treatments' are settings for various systematic or forcing parameters, and the response is the output of the climate model or epidemic model. Principles of experimental design are also widely used in clinical trials, where the experimental units are often patients, the treatments are medical interventions, and the response is some measure of efficacy of the treatment.

If the experimenter is able to ensure that the experimental units are homogeneous, and the treatments are assigned randomly, then there is some basis for attributing a difference in response under different treatments to the effect of the treatment; in some contexts the effect might be presumed then to be a causal effect. In most settings the randomization is subject to some constraints; for example experimental units might be subdivided into more homogeneous groups, conventionally called blocks, and treatments assigned to units at random within blocks. In clinical trials it is more or less impossible to ensure homogeneity of treatment groups, and several background variables will be recorded in order to attempt to assess whether an observed difference between two treatments might be ascribed to some other feature, such as, for example, a possibly small but important age difference between the groups. Randomization will on average balance out differences on all these so-called confounding variables, but with small groups of patients the balance may be far from perfect. In computer experiments such elaborate protections will not normally be needed, although it might be used if there could be, for example, some potential drift in conditions over time.

Table 1: A 2^4 factorial design of 16 runs, with the response labelled according to conventional notation for the factor levels.

run	A	B	C	D	response
1	-1	-1	-1	-1	$y_{(1)}$
2	-1	-1	-1	$+1$	y_d
3	-1	-1	$+1$	-1	y_c
4	-1	-1	$+1$	$+1$	y_{cd}
5	-1	$+1$	-1	-1	y_b
6	-1	$+1$	-1	$+1$	y_{bd}
7	-1	$+1$	$+1$	-1	y_{bc}
8	-1	$+1$	$+1$	$+1$	y_{bcd}
9	$+1$	-1	-1	-1	y_a
10	$+1$	-1	-1	$+1$	y_{ad}
11	$+1$	-1	$+1$	-1	y_{ac}
12	$+1$	-1	$+1$	$+1$	y_{acd}
13	$+1$	$+1$	-1	-1	y_{ab}
14	$+1$	$+1$	-1	$+1$	y_{abd}
15	$+1$	$+1$	$+1$	-1	y_{abc}
16	$+1$	$+1$	$+1$	$+1$	y_{abcd}

2 Factorial experiments

A very useful class of designed experiments are *factorial* experiments, in which the treatments are combinations of levels of several factors. These are used in many applications of experimental design, but especially in technological experiments, where the factors might be, for example, time, concentration, pressure, temperature, etc. It is very common to use a small number of levels for each of the factors, often just two levels, in which case a design with k treatment factors has 2^k treatments and is called a 2^k factorial design. As an example, in a computer experiment, if there were 10 systematic parameters then a full 2^{10} factorial might have each systematic parameter set at $\pm 1\sigma$; of course in this case it would be usual as well to have one or more runs at the central 'mean value' or 'best guess' of all the systematics.

A 2^k factorial design is to be contrasted with a one-factor-at-a-time, or OFAT, design, where, for example, a single simulation run would keep 9 of the 10 systematics at their mean values and use $+1\sigma$ for the 10th systematic; the next run would do the same but use -1σ for the 10th systematic, and subsequent runs would proceed through the other values. An OFAT design has the advantage that if a large change is observed in a single run, the change can be attributed to the systematic that was altered in that run, but it is a very inefficient way to extract this information. In fact the mean effects of each systematic can be estimated in a 2^k factorial design with considerable savings.

Table 1 gives the settings for a 2^4 factorial experiment; usually the order of the runs would be randomized, but the structure of the experiment is easier to see in the un-randomized form. The run called '1', for example, has all four factors set to their low level, whereas the run called '2', has factors A, B and C set to their low level and D set to its high level. Note that the estimated effect in going from the low level of A, say, to the high level of A, is based on comparing the averages of 8 observations taken at the low level with 8 observations taken at the high level. Each of these averages has a variance equal to $1/8$ the variance in a single observation, or in other words to get the same information from an OFAT design we would need 8 runs with A at 1σ and 8 runs with A at $+1\sigma$, all other factors held constant. Repeating this for each factor would require 64 runs, instead of 16. The balance of the 2^4 design ensures that we can estimate the effects for each of the four factors in turn from the average of 8 observations at

Table 2: The 2^4 factorial showing all of the interaction effects.

run	**A**	**B**	**C**	**D**	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
1	-1	-1	-1	-1	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	-1	-1	-1	-1	$+1$
2	-1	-1	-1	$+1$	$+1$	$+1$	-1	$+1$	-1	-1	-1	$+1$	$+1$	$+1$	-1
3	-1	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$	$+1$	-1
4	-1	-1	$+1$	$+1$	$+1$	-1	-1	-1	-1	$+1$	$+1$	$+1$	-1	-1	$+1$
5	-1	$+1$	-1	-1	-1	$+1$	$+1$	-1	-1	$+1$	$+1$	$+1$	-1	$+1$	-1
6	-1	$+1$	-1	$+1$	-1	$+1$	-1	-1	$+1$	-1	$+1$	-1	$+1$	-1	$+1$
7	-1	$+1$	$+1$	-1	-1	-1	$+1$	$+1$	-1	-1	-1	$+1$	$+1$	-1	$+1$
8	-1	$+1$	$+1$	$+1$	-1	-1	-1	$+1$	$+1$	$+1$	-1	-1	-1	$+1$	-1
9	$+1$	-1	-1	-1	-1	-1	-1	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	-1	-1
10	$+1$	-1	-1	$+1$	-1	-1	$+1$	$+1$	-1	-1	$+1$	-1	-1	$+1$	$+1$
11	$+1$	-1	$+1$	-1	-1	$+1$	-1	-1	$+1$	-1	-1	$+1$	-1	$+1$	$+1$
12	$+1$	-1	$+1$	$+1$	-1	$+1$	$+1$	-1	-1	$+1$	-1	-1	$+1$	-1	-1
13	$+1$	$+1$	-1	-1	$+1$	-1	-1	-1	-1	$+1$	-1	-1	$+1$	$+1$	$+1$
14	$+1$	$+1$	-1	$+1$	$+1$	-1	$+1$	-1	$+1$	-1	-1	$+1$	-1	-1	-1
15	$+1$	$+1$	$+1$	-1	$+1$	$+1$	-1	$+1$	-1	-1	$+1$	-1	-1	-1	-1
16	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$

the high level compared to 8 observations at the low level: for example the main effect of D is estimated by

$$(y_{abcd} - y_{abc} + y_{abd} - y_{ab} + y_{acd} - y_{ac} + y_{ad} - y_a + y_{bcd} - y_{bc} + y_{bd} - y_b + y_{cd} - y_c + y_d - y_{(1)})/8,$$

and similar estimates can be constructed for the effects of B and C.

Note that by constructing these four estimates we have used four linear combinations of our 16 observations. One linear combination, the simple average, is needed to set the overall level of response, leaving 11 linear combinations not yet used to estimate anything. These combinations are in fact used to estimate the *interactions* of various factors, and the full set of combinations is given by the set of signs in Table 2.

For example, the interaction of factors A and B is estimated by the contrast given by the fourth column of table 2:

$$\{y_{abcd} + y_{abc} + y_{abd} + y_{ab} - y_{bcd} - y_{bc} - y_{bd} - y_b - (y_{acd} + y_{ac} + y_{ad} + y_a - y_{cd} - y_c - y_d - y_{(1)})\}/8$$

which takes the difference of the difference between responses with A at high level and A at low level with the difference between responses with B at high level and B at low level. The column of signs in Table 2 for the interaction effect AB was obtained simply by multiplying the A column by the B column, and all the other columns are similarly constructed.

This illustrates two advantages of designed experiments: the analysis is very simple, based on linear contrasts of observations, and as well as efficiently estimating average effects of each factor, it is possible to estimate interaction effects with the same precision. Interaction effects can never be measured with OFAT designs, because two or more factors are never changed simultaneously.

The analysis, by focussing on averages, implicitly assumes that the responses are best compared by their mean and variance, which is typical of observations that follow a Gaussian distribution. However the models can be extended to more general settings, as will be briefly discussed in the next section.

Table 3: A screening design for 7 factors in 8 runs, built from a 2^3 factorial design.

run	A	B	C	D	E	F	G
1	-1	-1	-1	$+1$	$+1$	$+1$	-1
2	-1	-1	$+1$	-1	-1	$+1$	$+1$
3	-1	$+1$	-1	-1	$+1$	-1	$+1$
4	-1	$+1$	$+1$	$+1$	-1	-1	-1
5	$+1$	-1	-1	$+1$	-1	-1	$+1$
6	$+1$	-1	$+1$	-1	$+1$	-1	-1
7	$+1$	$+1$	-1	-1	-1	$+1$	-1
8	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$

In most applications the interpretation of 3- and 4- factor interactions would be rather difficult, and in fact these higher order interactions might be expected to be zero. If they are indeed zero, then 5 of the contrasts outlined in Table 2 are estimating zero, and their squares could then be pooled to provide an estimate of the variance of a single observation, with 4 degrees of freedom. This pooling of higher order interactions is often used in settings where the interactions are expected to be small, and no external estimate of variance is available. Sometimes even two-factor interactions are pooled and used to estimate the error.

Alternatively, we could assign new factors to the higher order interactions, leading to the class of fractional factorial designs. For example, we could use introduce a fifth factor, E, to the 2^4 factorial of Table 2, using the signs for the $ABCD$ interaction. That is, in the first run E would be set to its high level, in the second run to its low level, in the third run to its low level, and so on, following the pattern of $+1$ and -1 in the last column of Table 2. The resulting contrast $(y_{(1)} - y_a - y_b + y_{ab} \pm \dots)/8$ is estimating the main effect of factor E (i.e. the difference between responses on the high level of E to the low level of E), but it is also estimating the $ABCD$ interaction: these two effects are completely aliased. The working assumption is that the $ABCD$ interaction is likely to be very small, so any observed effect can be attributed to E. The main effects of A, B, C and D are estimated as before, and we now have information on 5 factors from a 16 run design. However all the main effects are aliased with 4 factor interactions: for example A is aliased with $BCDE$, B with $ACDE$, and so on. Further, all 2 factor interactions are aliased with 3 factor interactions. Again, the working assumption is typically that any observed effect is more likely to be due to a 2 factor interaction than a 3 factor interaction.

This process can be continued; we might for example assign a new factor F, say, to the ABC interaction (which is aliased with DE), giving a 2^{6-2} design, sometimes called a $1/4$ fraction of a 2^6. This allows us to assess the main effect of 6 factors in just 16 runs, instead of 64 runs, although now some 2 factor interactions will be aliased with each other.

There are very many variations on this idea; one is the notion of a screening design, in which only main effects can be estimated, and everything else is aliased. The goal is to quickly assess which of the factors is likely to be important, as a step in further experimentation involving these factors and their interactions. Table 3 shows an 8 run screening design for 7 factors. The basic design is the 2^3 factorial in factors A, B and C shown in the first 3 columns; then 4 new factors have been assigned to the columns that would normally correspond to the interactions BC, AC, AB and ABC.

There is a very large literature on fractional factorial designs; a good introduction aimed at physicists is given in [1] and much of this paper draws on those ideas. A detailed but quite accessible introduction is given in [2]. Some advantages of these fractional factorial designs is the ability to screen a large number of factors in a few runs, in settings where many factors are expected to be inactive. More

Table 4: Data and design for a 2^{5-1} factorial.

A	B	C	D	E	response
-1	-1	-1	-1	$+1$	29.17
-1	-1	-1	$+1$	-1	29.39
-1	-1	$+1$	-1	-1	22.13
-1	-1	$+1$	$+1$	$+1$	27.64
-1	$+1$	-1	-1	-1	11.53
-1	$+1$	-1	$+1$	$+1$	16.20
-1	$+1$	$+1$	-1	$+1$	14.99
-1	$+1$	$+1$	$+1$	-1	19.29
$+1$	-1	-1	-1	-1	16.30
$+1$	-1	-1	$+1$	$+1$	22.40
$+1$	-1	$+1$	-1	$+1$	19.42
$+1$	-1	$+1$	$+1$	-1	23.85
$+1$	$+1$	-1	-1	$+1$	6.70
$+1$	$+1$	-1	$+1$	-1	13.17
$+1$	$+1$	$+1$	-1	-1	8.53
$+1$	$+1$	$+1$	$+1$	$+1$	19.04

complete fractional factorials, such as $1/2$ fractions or $1/4$ fractions permit assessing a small number of main effects and two-factor interactions. Often a number of the inactive effects can be pooled to provide an internal estimate of variability.

These designs are more complicated to run than OFAT designs, as several factors settings need to be changed with each run. If some factor levels are difficult to change, for example temperature of an oven, in a manufacturing context, then a full factorial design will not be feasible. In such cases it is often possible to have an 'outer' factorial with the difficult-to-change factors, and an 'inner' factorial of the other factors; the analysis of these *split plot* designs is a little more complex. There is a lot of information in a single run of a factorial design, so if a run is lost, the associated balance is lost along with quite a bit of information. It is often necessary to block runs to ensure homogeneity; for example if all runs cannot be completed in a single day and there is concern about changes in conditions from one day to the next. This is relatively straightforward to implement but the analysis of the results is again a little more complicated.

3 Analysis of the data

Implicit in the discussion above is a linear model with Gaussian error

$$y = Z\beta + \epsilon$$

where y is an $n \times 1$ vector of responses, and Z is the so-called *design* matrix, with n rows and p, say, columns, and we assume ϵ follows a Gaussian distribution with mean 0 and covariance matrix σ^2 times the identity. This is exactly a linear regression formulation, but the design matrix Z has a particularly simple form. The first column is a column of $+1$, and the remaining columns have elements ± 1 according to the factorial structure. Table 2 gives an example: in a single run of a 2^4 factorial, y will have length 16, and the 16×16 matrix Z has columns 2 through 16 given by the columns of this table. Any standard regression package will fit this model, although there will be no degrees of freedom available to estimate the error. By specifying a simpler model with just main effects and 2-factor interactions, so that Z now

has dimension 16×11, we will have 5 degrees of freedom left to estimate σ^2. The design matrix Z is completely orthogonal, which makes the least squares fitting of the model particularly simple; in fact it can be computed by hand, and an early algorithm to do this computation invented by Yates is a pre-cursor to the fast Fourier transform.

Most statistical software can deduce from the specification of the model which effects are aliased, and in some packages, including `R` and `Splus` it is relatively easy to produce a graphical display of the estimated effects that allows one to assess which effects are non-zero, at least in part so that the 'nearly zero' effects can be pooled to estimate the error.

An example of the standard linear analysis for a 2^{5-1} factorial, carried out in `R` is given in Figure 1. Figure 2 gives the display of estimated effects described above. It is conventional to ignore the sign of the effects, so the ordered (absolute) values are then plotted against the expected values of ordered (absolute) standard Gaussian variables. The data and design are given in Table 4, following the 2^4 design of Table 1, but assigning factor E to the 4-factor interaction.

If the response is non-Gaussian, then the model will normally assume that some transformation of the mean of the response follows a linear structure of the form $Z\beta$; these models are often called generalized linear models in the statistical literature. For example if y follows a Poisson distribution with mean μ, we might assume $\log \mu = Z\beta$, and fit the model by maximum likelihood. If y is a proportion, then a version of logistic regression is often used, assuming that $\log\{p/(1-p)\} = Z\beta$, where p is the mean value of y. These models can be fit using the `glm` command of `R`. An example of a fractional factorial fit to Poisson data is given in [3], §5.4. An alternative is to transform the responses to something approximately Gaussian, and use the linear model formulation above. If the response is more complex, such as a histogram, then analysis might proceed by constructing one or more derived responses, at least as a first step.

4 Response surface designs

Very often, especially in manufacturing settings, the factors correspond to underlying quantitative variables, and the levels, denoted ± 1 in the previous section, are codes for particular numerical values: temperatures at 80 and 100 degrees, for example. In such cases the choice of factor levels involves both subject matter expertise, and at least in the early stages, considerable guesswork. As well, the goal of the experiment might not be to compare responses at different factor settings, but to find the combination of factor settings that leads to minimum or maximum response.

Factorial designs adapted to these conditions are called response surface designs. The basic idea is that the response y is a continuous function of some input variables x_1, x_2, and so on, and factorial designs are used sequentially to explore the shape of the response surface. Sequential experimentation in the relevant range of x-space usually begins with a screening design, to quickly assess which of several factors have the largest effect on the response. Then second stage is a factorial design at new values of the underlying variables, chosen in the direction of increasing (or decreasing) response. Near the maximum additional factor levels are added, to model curvature in the response surface. In the setting of simulation experiments, the goal might be to see which values of the systematics produce simulated data consistent with the observations; thus we would be seeking to minimize a derived response measuring discrepancy of the simulation with the data would be. Another goal might be simply to see which systematic parameters affect the simulation output, and whether they affect it linearly or in a more complex fashion.

Figure 3 is adapted from [4], although similar pictures can be found in, for example [2], and other treatments of response surface methods, such as [5]. The contours of a smooth response surface in two quantitative variables are indicated, along with an initial 2^2 factorial design. The $+$ symbols indicate the design points for the first experiment, and the results could lead to a second 2^2 factorial carried out at the circles indicated. The results would show that the maximum is 'surrounded', so to speak, at which

```
>  y <- c(29.17,29.39,22.13,27.64,11.53,16.20,14.99,19.29,16.30,
> +   22.40,19.42,28.85,6.70,13.17,8.53,19.04)
> A <- c(rep(-1,8),rep(1,8))
> B <- c(rep(-1,4),rep(1,4),rep(-1,4),rep(1,4))
> C <- rep(c(rep(-1,2),rep(1,2)),4)
> D <- rep(c(-1,1),8)
> E <- A*B*C*D
> A <- factor(A); B <- factor(B); C <- factor(C); D <- factor(D); E <- factor(E)
> # The factor() tells R to interpret the levels of A as qualitative instead of quantitative.
> # This enables use of abbreviated notation for the  model, as shown below.

> fact.lm <- lm(y~A*B*C*D*E) # lm is the general linear model fitting routine
# we avoided specifying the full matrix Z, although this could have been done instead

> fact.lm                 # there are many ways to summarize the output
Call:
lm(formula = y ~ A * B * C * D * E)

Coefficients:
   (Intercept)              A1              B1              C1
        25.686          -9.386         -14.156          -3.556
            D1              E1           A1:B1           A1:C1
         3.704           3.484           1.697           4.878
         B1:C1           A1:D1           B1:D1           C1:D1
         3.367           4.453           1.173           3.073
         A1:E1           B1:E1           C1:E1           D1:E1
        -1.238           0.612          -0.448          -4.303
      A1:B1:C1        A1:B1:D1        A1:C1:D1        B1:C1:D1
            NA              NA              NA              NA
  # ...[other NAs deleted]

> coef(fact.lm)[2:16]
     A1      B1      C1      D1      E1   A1:B1   A1:C1   B1:C1   A1:D1
 -9.386 -14.156  -3.556   3.704   3.484   1.697   4.878   3.367   4.453
  B1:D1   C1:D1   A1:E1   B1:E1   C1:E1   D1:E1
  1.173   3.073  -1.238   0.612  -0.448  -4.303

> library(faraway) # a package named "faraway" gives easy access to half normal plots
> halfnorm(coef(fact.lm)[2:16],labs=fact.names,main="Half-normal plot for identifying
 effects", ylab="effect estimates")

> # A simpler model with just factors A, B and C and their interactions
> anova(lm(y~A+B+C+A:B+A:C+B:C+A:B:C))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
A          1     81      81    3.70 0.0907 .
B          1    461     461   21.11 0.0018 **
C          1     14      14    0.65 0.4444
A:B        1      3       3    0.13 0.7257
A:C        1     24      24    1.09 0.3270
B:C        1     11      11    0.52 0.4915
A:B:C      1     19      19    0.85 0.3840
Residuals  8    175      22
```

Fig. 1: Some R code illustrating analysis of a factorial design

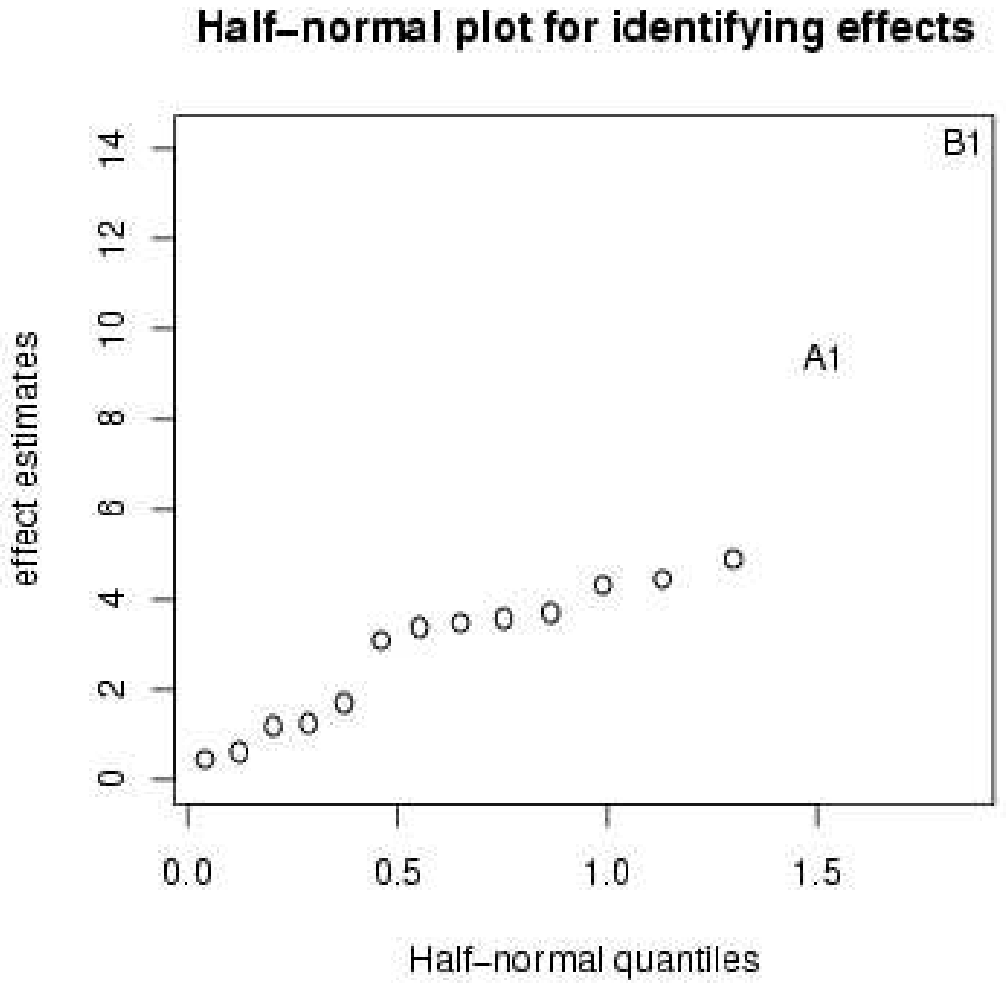

Fig. 2: A graphical display of the estimated effects for the data from Table 4; the two largest effects are labelled.

stage it would be usual to add center points and radial points to attempt to quantify the curvature at the maximum. Note that with an OFAT design, the sequential stages of experimentation could only proceed along lines parallel to the coordinate axes, which is less efficient unless the axes of the elliptical contours are aligned with the coordinate axes.

A two-level factorial design can only detect linear effects of x_1 and x_2, and their interaction, x_1x_2. The other quadratic effects, x_1^2 and x_2^2 need a minimum of three levels to be estimated. A very common approach to estimating a smooth, curved, response surface is to add center points at $(0, 0)$, often replicated, to give an internal estimate of error, and then to add further points on the radius of a circle. Such designs are called *central composite* designs. This is illustrated for two factors in Figure 4, but the idea is very general.

5 More specialized designs

The 8 run screening design illustrated in Table 3 is a 2^{7-4} fractional factorial, but is also an example of an *orthogonal array*, which is by definition an array of symbols, in this case ± 1, in which every symbol appears an equal number of times in each column, and any pair of symbols appears an equal number of times in any pair of columns. An orthogonal array of size $n \times (n-1)$ with two symbols in each column

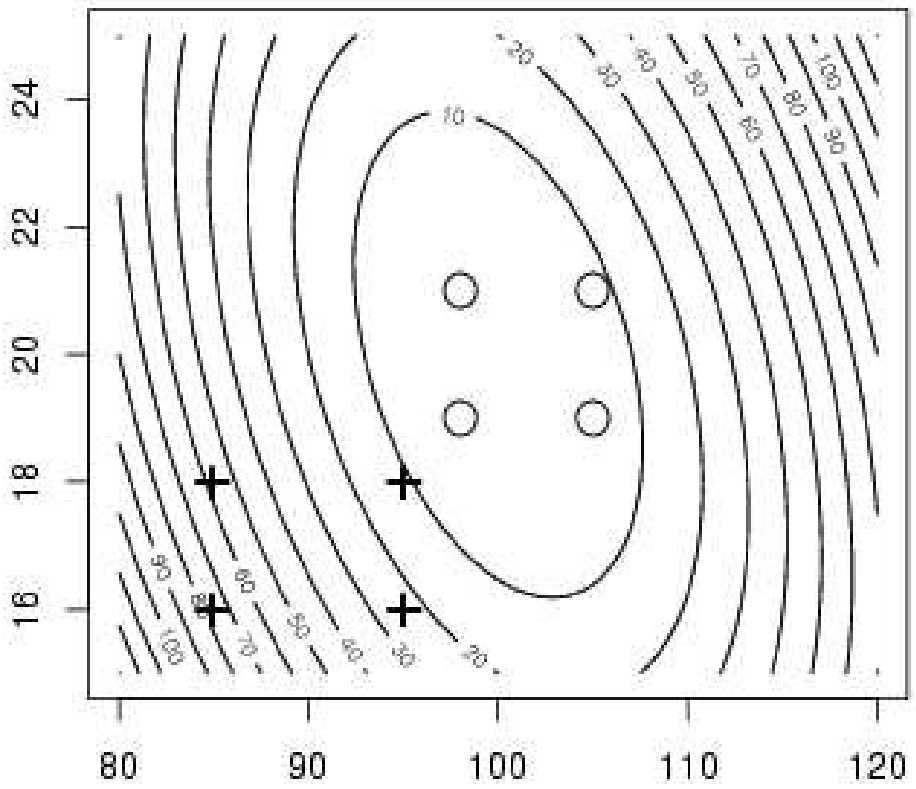

Fig. 3: Two 2^2 experiments to explore a response surface: + shows the design points for the first experiment, and o shows the design points for the second.

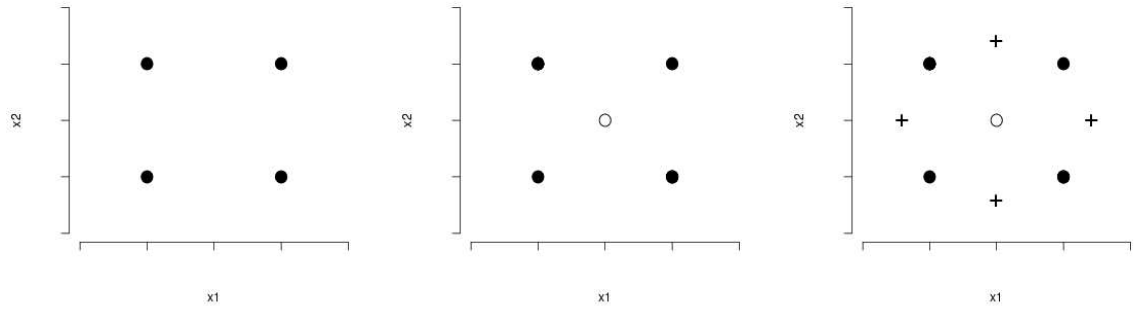

Fig. 4: A series of experiments adding points to capture (to first order) the nonlinearity in the response surface; the third is a central composite design.

Table 5: An orthogonal array for 6 factors each at 3 levels

run	A	B	C	D	E	F
1	-1	-1	-1	-1	-1	-1
2	-1	0	0	0	0	0
3	-1	$+1$	$+1$	$+1$	$+1$	$+1$
4	0	-1	-1	0	0	$+1$
5	0	0	0	$+1$	$+1$	-1
6	0	$+1$	$+1$	-1	-1	0
7	$+1$	-1	0	-1	$+1$	0
8	$+1$	$+1$	-1	$+1$	0	-1
9	$+1$	$+1$	-1	$+1$	0	$+1$
10	-1	-1	$+1$	$+1$	0	0
11	-1	0	-1	-1	$+1$	$+1$
12	-1	$+1$	0	0	-1	-1
13	0	-1	0	$+1$	-1	$+1$
14	0	0	$+1$	-1	0	-1
15	0	$+1$	-1	0	$+1$	0
16	$+1$	-1	$+1$	0	$+1$	-1
17	$+1$	0	-1	$+1$	-1	0
18	$+1$	$+1$	0	-1	0	$+1$

specifies an n-run screening design for $n-1$ factors. The designs with symbols ± 1 are called Plackett-Burman designs and Hadamard matrices defining them have been shown to exist for all multiples of four up to 424. A Plackett-Burman design is used for studying simulations in [9].

More generally, an $n \times k$ array with m_i symbols in the ith column is an *orthogonal array* of strength r if all possible combinations of symbols appear equally often in any r columns. The symbols correspond to levels of a factor. Table 5 gives an orthogonal array of 18 runs, for 6 factors with three levels each.

Orthogonal arrays are particularly popular in applications of so-called Taguchi methods in technological experiments and manufacturing. An extensive discussion is given in [6]. In the discussion of the talk at CERN, Jim Linneman pointed out that these applications could be relevant to HEP experiments at the stage at which the equipment is being designed, tested and manufactured.

A related approach to the exploration of possibly complex response surfaces is the use of *space filling* designs. These are especially popular in computer experiments and simulation experiments, as well as in numerical integration, where the method is known as quasi Monte Carlo. In the approximation of a multi-dimensional integral

$$\int_{R^k} f(x)dx \approx \frac{1}{n}\sum f(X_i)$$

the X_i are taken to be 'space-filling' points, whereas in simple Monte Carlo the X_i would be sampled randomly, possibly using a uniform distribution on each coordinate. The difficulty with the simple approach is that in high dimensions the compounding of the uniform points tends to concentrate on the outer shell, and interior points are too rarely sampled. Orthogonal arrays can be used as the basis of space-filling designs, and in this application are often called Latin hypercube designs. A good reference is [7]; however in the discussion of the talk at CERN Fred James said that in fact he had little success with space-filling designs in high dimensional integrations.

6 An example motivated by miniBoone

In this section I report on some preliminary work by Zi Jin, Radford Neal and myself. This does not in fact use the orthogonal constructions described in the preceding sections, but is some preliminary work to see if these methods could provide improvement in simulation experiments. The basic ideas were described in the context of the miniBoone experiments by Byron Roe at the Banff workshop in summer 2006.

Suppose that a simulation run generates M background events and mistakenly identifies y of these a signal, with some small probability p, say $p = 0.001$ or $p = 0.0001$. We use as a first approximation the Poisson model for y with mean p. This mean is assumed to depend on various settings for the systematics, presumably in a fairly complex way that could be explored using factorial designs and other concepts from the previous sections.

We bypass that by making the very rough approximation that p depends on these systematics via a Gamma distribution with parameters a and b. The mean and variance of the gamma distribution are a/b and a/b^2, respectively, so a and b would be chosen to allow p to vary between, say $\pm 3\sigma$ about its mean. We then explore how the Fisher information in a full set of N simulation runs depends on the trade-off between sampling K different values of these systematics or sampling M events at each value of the systematics, under the constraint $N = MK$.

For example, suppose p is approximately 0.0001. If we fix the total number of simulations N at $2,000,000$, then the optimal split is roughly $M = 160,000$ runs at each of $K = 13$ different parameter settings. This represents a 10-fold increase in precision (inverse variance) over the rather arbitrary choice of $M = 100,000$ and $K = 20$. The improvement indicated by these preliminary results suggests that it will be worthwhile to investigate the information in the more complex orthogonal array based designs. A more complete account of these results will be presented elsewhere.

7 A short guide to the references

Much of this paper was inspired by two articles by Gunter, [1] and [4], which are written for scientists. The book by Box, Hunter and Hunter, [2] has very detailed explanation of the sequential nature of experimentation in industrial applications, and the book by Montgomery [5] has a more concise discussion, in something of a cookbook style. A somewhat more theoretical approach is given in [3]. Ref. [6] has a lot of material on orthogonal arrays, but their use in numerical integration is probably best approached from [7]. Ref [8] is an introduction to the general field of computer experiments, but [9] describes applications that are probably closer to those needed for HEP simulations.

References

[1] B.H. Gunter, *Computers in Physics* **7**, 262–272.

[2] Box, G.E.P., Hunter, J.S. and Hunter, W. G. (1978) *Statistics for Experimenters.* Wiley & Sons, New York.

[3] Cox, D.R. and Reid, N. (2000) *The Theory of Design of Experiments.* Chapman & Hall/CRC, London.

[4] Gunter, B.H. (2007) "Sequential Experimentation", to appear in *Encyclopedia of Statistics in Quality and Reliability*, Wiley & Sons, New York.

[5] Montgomery, D. C. (2005). *Design and Analysis of Experiments (6th ed.)* Wiley & Sons, New York.

[6] Wu, C.F.J. and Hamada, M. *Experiments: Planning, Analysis, and Parameter Design Optimization.* Wiley & Sons, New York.

[7] Owen, A. (1992). Orthogonal arrays for computer experiments, integration and visualisation. *Statistica Sinica* **2**, 459–462.

[8] Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989). Design and analysis of computer experiments. *Statistical Science* **4** 409–423.

[9] Welsh, J.P., Koenig, G.G. and Bruce, D. (1997). Screening design for model sensitivity studies. *J. Geophys. Res.* **102** D14, 16,499–16,505.